Imputation of Assay Bioactivity Data using Deep Learning

T.M. Whitehead,[*,†] B.W.J. Irwin,[‡] P. Hunt,[‡] M.D. Segall,[‡] and G.J. Conduit[¶]

†*Intellegens, Eagle Labs, Chesterton Road, Cambridge, CB4 3AZ, United Kingdom*
‡*Optibrium, F5-6 Blenheim House, Cambridge Innovation Park, Denny End Road, Cambridge, CB25 9PB, United Kingdom*
¶*Cavendish Laboratory, University of Cambridge, J.J. Thomson Avenue, Cambridge, CB3 0HE, United Kingdom*

E-mail: tom@intellegens.ai

Abstract

We describe a novel deep learning neural network method and its application to impute assay pIC$_{50}$ values. Unlike conventional machine learning approaches, this method is trained on sparse bioactivity data as input, typical of that found in public and commercial databases, enabling it to learn directly from correlations between activities measured in different assays. In two case studies on public domain data sets we show that the neural network method outperforms traditional quantitative structure-activity relationship (QSAR) models and other leading approaches. Furthermore, by focussing on only the most confident predictions the accuracy is increased to $R^2 > 0.9$ using our method, as compared to $R^2 = 0.44$ when reporting all predictions.

1 Introduction

Accurate compound bioactivity and property data are the foundations of decisions on the selection of hits as the starting point for discovery projects, or the progression of compounds through hit to lead and lead optimisation to candidate selection. However, in practice, the experimental data available on potential compounds of interest are sparse. High-throughput screens may be run on a large screening collections, but these are costly, and thus applied infrequently, and the throughput of an assay usually comes with a trade-off against the quality of the measured data. As discovery projects progress and new compounds are synthesised, the increasing cost of generating high-quality data means that only the most promising compounds are advanced to these late-stage studies.

If one considers all of the compounds in a large pharmaceutical company's corporate collection and the assay endpoints that have been measured, only a small fraction of the possible compound-assay combinations have been measured in practice. Public domain databases are also sparsely populated; for example, the ChEMBL[1,2] data set is just 0.05% compete.

The implication of this is that a vast trove of information would be revealed if only a small fraction of these missing data could be filled in with high-quality results in a cost-effective way. New hits for projects targeting existing biological targets of interest and high-quality compounds, overlooked during optimisation projects, could be identified. Furthermore, compounds with results from early assays could be selected for progression with greater confidence if downstream results could be accurately predicted.

A common approach for prediction of compound bioactivities is the development of quantitative structure-activity relationship (QSAR) models.[3] These are generated using existing data to identify correlations between easily calculated characteristics of compound structures, known as descriptors, and their biological activities or properties. The resulting models can then be applied to new compounds that have not yet been experimentally tested, to predict the outcome of the corresponding assays. A wide range of statistical methods have been applied to build QSAR models, from simple linear regression methods such as partial least squares[4] to more sophisticated machine learning approaches such as random forests (RF),[5–9] support vector machines[10] and Gaussian processes.[11] Another approach is the profile-QSAR (pQSAR) method proposed by Martin *et al.*,[12,13] which uses a hierarchical approach to build a model of a bioactivity by using as inputs the predictions from QSAR models of multiple bioactivities that may be correlated. Recently, the application of advances in deep learning have been explored for generation of QSAR models;[14] while small improvements in the accurcay of predictions have been found, these methods have not generally resulted in a qualitative step forward for activity predictions.[15–17] One advantage of deep learning methods is the ability to train models against multiple endpoints simultaneously, so-called multi-target prediction. This enables the model to 'learn' where a descriptor correlates with multiple endpoints and hence improve the accuracy for all of the corresponding endpoints.

However, the sparse experimental data could reveal more information regarding the correlations between the endpoints of interest, if these could be used as *inputs* to a predictive model. Conventional machine learning methods cannot use this information as inputs because the bioactivity data are often incomplete, and so cannot be relied on as input. In this paper we present a novel deep learning framework, previously applied to materials discovery,[18–20] that can learn from and exploit information that is sometimes missing, unlike other contemporary machine learning methods. A further benefit of the proposed method is that it can estimate the uncertainty in each individual prediction, allowing it to improve the quality of predictions by focussing on only the most confident results.

We will compare the performance of our method to impute bioactivities with a RF, a commonly applied and robust QSAR machine learning method, a modern multi-target deep learning method, a leading matrix factorisation approach, and the second-generation pQSAR 2.0 technique.[13]

In Section 2 we present the underlying deep learning methodology to handle missing data and estimate uncertainty, along with details of the data sets used in this study, the accuracy metric, and other machine learning methods applied for comparison. Then in Section 3 we present two examples to assess the performance of the algorithm against current methods. Finally, in Section 4 we discuss our findings and potential applications of the results.

2 Methodology

The goal for the neural network tool is to predict and impute assay bioactivity values, by learning both the correlations between chemical descriptors and assay bioactivity values and also the correlations between the assay bioactivities. In Subsection 2.1 we introduce the data sets used to validate the approach, before turning in the following subsections to the description of the neural network method itself.

2.1 Data sets

Two data sets were used to train and validate the models: a set containing activities derived from five adrenergic receptor assays (hereafter described as the "Adrenergic set") and a data set comprised of results from 159 kinase assays proposed by Martin *et al.* as a challenging benchmark for machine learning methods[13] (the "Kinase set"). These data sets are summarised in Table 1. All of the data were sourced from binding assays reported in the ChEMBL database[1,2] and the assay data represented as pIC_{50} values (the negative log of the IC_{50} in molar units). In the case of the Adrenergic set, measurements from different assays were combined for each target activity and, where multiple values were available for the same compound, the highest pIC_{50} value was used, representing a 'worst case' scenario for selectivity. In the case of the Kinase data set, each activity was derived from a single assay, as defined in ChEMBL.

Table 1: A summary of the data sets used in the examples presented herein. The table shows the data set, the number of compounds and assays each contains, and the proportion of the compound-assay values that are filled.

Data set	Compounds	Assays	Filled
Adrenergic	1731	5	37.5%
Kinase	13998	159	6.3%

320 molecular descriptors were used to characterise the compounds in the data sets. These comprised whole-molecule properties, such as the calculated octanol:water partition coefficient (logP), molecular weight, topological polar surface area[21] and McGowan volume,[22] as well as counts of substructural fragments represented as SMARTS patterns.[23]

In the case of the Adrenergic set, we employed a five-fold cross-validation approach for building models and assessing their resulting accuracy. The compounds in the data set were randomly split into five disjoint subsets of equal size, the models were trained using four of the subsets, and then their accuracy evaluated on the remaining subset. We repreated this process using each of the subsets for testing, so that each compound was used as a test case for the tool. The Adrenergic data set is provided with the supporting information for this paper.

The Kinase set was provided in the supporting information of the paper by Martin *et al.*[13] as a challenging benchmark for machine learning methods. In this case, the data set was split by Martin *et al.* into independent training and test sets. The data were initially clustered for each assay and the members of the clusters used as the training set, leaving the outliers from this clustering procedure as the data against which the resulting models were tested. This procedure means that the test data is not representative of the data used to train the models, making this a difficult test of a machine learning method's ability to extrapolate outside of the chemical space on which it was trained. Martin *et al.* described this as a 'realistic' test set, designed to be more representative of real working practices in an active chemistry optimisation project, where new compounds are continuously proposed that extend beyond the chemical space that has previously been explored. Because the clustering was carried out on a per-assay basis, some compounds appear in both the train and test sets: but the assay data for each compound is split between the sets, so that none of the same assay/compound pairs appear in both the train and test set and the validation is against a robust, disjoint test case. The Kinase data set is provided with the supporting information for this paper.

2.2 Performance Metric

To assess the performance of the models we use the coefficient of determination R^2 for each assay in the test set:

$$R^2 = 1 - \frac{\sum_i (y_i^{pred} - y_i^{obs})^2}{\sum_i (y_i^{obs} - \overline{y^{obs}})^2},$$

where y_i^{obs} is the i^{th} observed assay value and y_i^{pred} is the corresponding prediction. This is a more stringent test than the commonly used squared Pearson correlation coefficient, which is a measure of the fit to the best fit line between the predicted and observed values, while the coefficient of determination is a measue of the fit to the perfect identity line $y_i^{pred} = y_i^{obs}$. By definition, the coefficient of determination is less than or equal to the squared Pearson correlation coefficient.

For each of the methods, we report the mean of the R^2 across all of the assays in the test set to give an overall value.

2.3 Neural network formalism

We now turn to the neural network formalism. This algorithm is able to automatically identify the link between assay bioactivity values, and use the bioactivity data of other compounds to guide the extrapolation of the model, as well as using molecular descriptors as design variables. Furthermore, the method can estimate uncertainties in its predictions. The neural network builds on the formalism used to design nickel-base superalloys, molybdenum alloys, and identify erroneous entries in materials databases.[18–20] We describe here the core neural network and the first novel aspect, the ability to estimate the uncertainty in the predictions, before Section 2.4 details the second novel part of the algorithm: how to handle missing data, necessary to capture bioactivity-bioactivity correlations.

Each input vector $\mathbf{x} = (x_1, \ldots, x_{A+D})$ to the neural network contains values for $D = 320$ molecular descriptors and $A = 5$ (for the Adrenergic data set) or $A = 159$ (for the Kinase data set) bioactivity values. The ordering of the elements of the input is the same for each compound, but otherwise unimportant. The output $(y_1, \ldots, y_{A+D})$ of the neural network consists of the original descriptors and the predicted bioactivities: only the elements $(y_1, \ldots, y_A)$ corresponding to predicted bioactivities are used for evaluating the network accuracy.

The neural network itself is a linear superposition of hyperbolic tangents

$$\mathbf{f} : (x_1, \ldots, x_i, \ldots, x_{A+D}) \mapsto (y_1, \ldots, y_j, \ldots, y_{A+D})$$

$$\text{with} \quad y_j = \sum_{h=1}^{H} C_{hj}\eta_{hj} + D_j,$$

$$\text{and} \quad \eta_{hj} = \tanh\left(\sum_{i=1}^{I} A_{ihj}x_i + B_{hj}\right) .$$

This neural network has a single layer of hidden nodes η_{hj} with parameters $\{A_{ihj}, B_{hj}, C_{hj}, D_j\}$ as shown in Figure 1. Each property y_j for $1 \leq j \leq A$ is predicted separately. We set $A_{jhj} = 0$ so the network will predict y_j without knowledge of x_j. Typically around five hidden nodes η_{hj} per output variable gives the best-fitting neural network. We use hyperbolic tangent activation functions to constrain the magnitude of η_{hj}, giving the weights C_{hj} sole re-

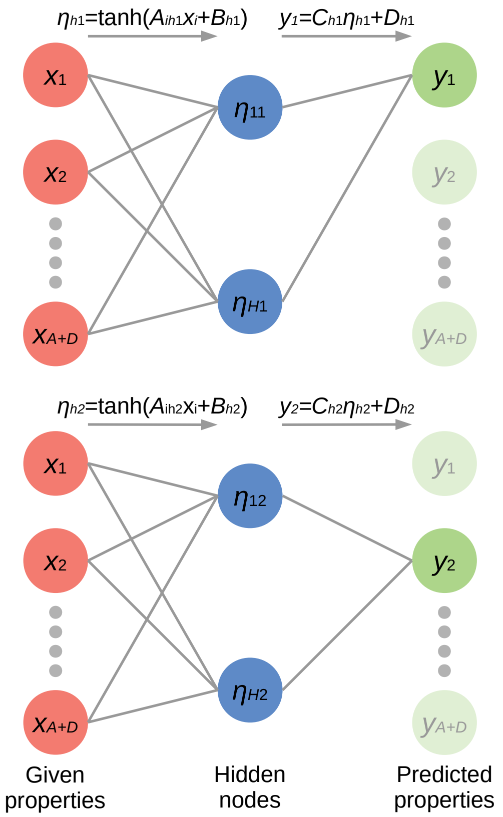

Figure 1: The neural network. The graphs show how the outputs for y_1 (top) and y_2 (bottom) are computed from all the inputs; similar graphs can be drawn for all other y_j to compute all the predicted properties. A linear combination (gray lines) of the given properties (red) are taken by the hidden nodes (blue), a non-linear tanh function, and a linear combination (gray lines) gives the predicted property (green).

sponsibility for the amplitude of the output response. Twelve separate networks were trained on the data with different weights[18–20] and their variance taken to indicate the uncertainty in the predictions accounting for both experimental uncertainty in the underlying data and the uncertainty in the extrapolation of the training data.[24,25] This is conceptually similar to the approach taken to uncertainty estimation in ensemble models, although here the underlying model is a deep neural network and the uncertainty estimates generated accurately represent the observed errors in the predictions, including uncertainty due to extrapolation that is poorly captured by random forest (see also Section 3.2).

2.4 Handling incomplete data

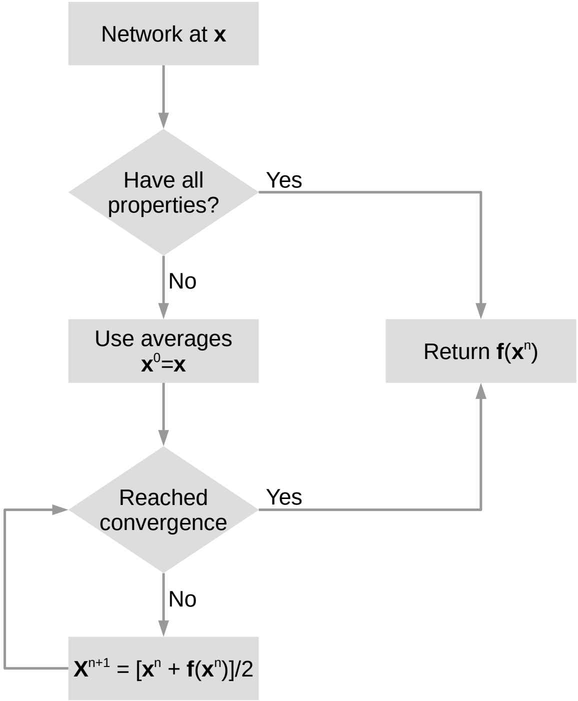

Figure 2: The data imputation algorithm for the vector $\mathbf{x}$ of the molecular descriptors and bioactivity values that has missing entries. We set $\mathbf{x}^0 = \mathbf{x}$, replacing all missing entries by averages across each assay, and then iteratively compute $\mathbf{x}^{n+1}$ as a function of $\mathbf{x}^n$ and $\mathbf{f}(\mathbf{x}^n)$ until we reach convergence after n iterations.

Experimental data are often incomplete – bioactivity values are not known for every compound and assay, and moreover the set of missing bioactivities is different for each compound. However, there is information embedded within bioactivity-bioactivity correlations. A typical neural network formalism requires that each property is either an input or output of the network, and all inputs must be provided to obtain a valid output. In contrast, we treat both the molecular descriptors and also the assay bioactivities as both inputs and outputs of the neural network and adopt an expectation-maximization algorithm,[26] where we first provide an estimate for the missing data, and use the neural network to iteratively improve that initial estimate.

The algorithm is shown in Figure 2. For any unknown bioactivities we first set missing values to the average of the bioactivity values present in the data set for that assay. With estimates for all values of the neural network we can then iteratively compute

$$\mathbf{x}^{n+1} = \frac{\mathbf{x}^n + \mathbf{f}(\mathbf{x}^n)}{2} \, .$$

The final predictions $(y_1, \ldots, y_A)$ are then the elements of this converged algorithm corresponding to the assay bioactivity predictions. The softening of the results by combining them with the existing predictions serves to prevent oscillations of the predictions, similar to the use of "shortcut connections" in ResNet.[27] Typically up to 5 iteration cycles were used to impute missing bioactivity values, using the same function $\mathbf{f}$ (as defined in Section 2.3) in every cycle. After 5 cycles the coefficient of determination R^2 in training improved by less than 0.01, comparable to the accuracy of the approach, confirming that we had used sufficient iteration cycles to reach convergence.

The parameters $\{A_{ihj}, B_{hj}, C_{hj}, D_j\}$ in the function $\mathbf{f}$ are then trained using simulated annealing[28] to minimize the least-square error of the predicted bioactivities $(y_1, \ldots, y_A)$ against the training data. At least 10^5 training rounds were used to reach convergence.

Hyperparameters, in particular the number of hidden nodes per output, the number of iteration cycles, and the number of training rounds, were selected using random holdout validation on each training data set, without reference to the corresponding test set.

2.5 Other machine learning methods

We compare our neural network algorithm with a variety of other popular machine learning approaches from the literature. RF methods[5–9] are a popular method of QSAR analysis, building an ensemble of decision trees to predict individual assay results. Because decision trees require all their input data to be present when they are trained, it is not possible to build RF models using sparse bioactivity data as input, and RF must rely purely on chemical descriptors. We used the scikit-learn[29] implementation of the regression RF method.

For a comparison with a modern deep learning approach, we also built a conventional multi-target deep neural network (DNN) model[30] using TensorFlow.[31] The model took linear combinations of descriptors as inputs, with eight fully connected hidden layers with 512 hidden nodes, and output nodes that gave the predicted assay results. The ELU activation function was used for all layers, and the network was trained using Adam backpropagation with Nesterov momentum[32] and a masked loss function to handle missing values. A principal component analysis (PCA) was performed on the descriptors to select the subset of linear combinations of descriptors that captured 90% of the variance across the full descriptor set to avoid overfitting of the DNN through the use of too many descriptors.

A popular method of analysing sparse databases is matrix factorisation,[33] where the matrix of compound-assay bioactivity values is approximately factorised into two lower-rank matrices that are then used to predict bioactivity values for new compounds. Matrix factorisation was popularised through its inclusion in the winning entry of the 2009 Netflix Prize.[34] We used the modern Collective Matrix Factorisation (CMF)[35,36] implementation of matrix factorisation, which makes effective use of the available chemical descriptors as well as bioactivity data, with separate latent features specialising in handling the descriptors.

We also compare to the profile-QSAR 2.0 method of Martin *et al.*,[13] which builds a linear partial least squares (PLS) model of assay bioactivities from the predictions of random forest models for each assay individually. In the 2.0 version of the profile-QSAR method the RF predictions for an assay are not used as input to the PLS model for that assay.

3 Imputing assay bioactivities

We present two tests of the performance of the deep learning formalism to impute assay bioactivity values. In each case we use disjoint training and validation data to obtain a true statistical measure, the coefficient of determination, for the quality of the trained models.

3.1 Adrenergic receptors

We first present a case study using the Adrenergic data set described in Section 2.1. We train two classes of model: the first uses complete compound descriptor information to predict the bioactivity values, and the second class uses both the chemical descriptors and also the bioactivity-bioactivity correlations.

We first train a neural network to take only chemical descriptors and predict assay bioactivities. This approach is similar to traditional QSAR approaches, although it offers the advantage of being able to indirectly learn the relationships between assay bioactivities through the iterative cycle described in Figure 2. We train the neural network providing as input the N descriptors with the highest average absolute Pearson correlation against the five targets, with N varying between 0 and the full set of 320 descriptors. The grey line in Figure 3 shows that the neural network, predicting based purely on descriptors, achieves a peak $R^2 = 0.60 \pm 0.03$ against the assays when using 50 descriptors: fewer descriptors do not provide a sufficient basis set, whereas more descriptors

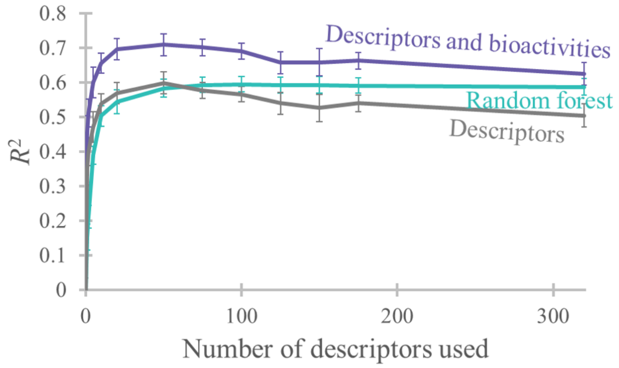

Figure 3: The coefficient of determination for predicting the activity of the adrenergic receptors with number of chemical descriptors. The magenta line is when the neural network is trained with both the activities and descriptors present, the grey line with just the descriptors, and the cyan line is for random forest. Error bars represent the standard error in the mean R^2 over five-fold cross-validation.

over-fit the data. The neural network not requiring the full set of chemical descriptors to provide a high-quality fit enables us to focus attention on the key descriptors, and hence chemical features, that influence bioactivity against these targets.[37] We compare the neural network result to traditional random forest, using the same descriptor sets, which achieves a similar value of $R^2 = 0.59 \pm 0.02$ using 100 descriptors.

We next train a fresh neural network but include the possibility of bioactivity-bioactivity correlations. With a total of 5 assays, this allows up to 4 additional input values per target as bioactivity values for every other assay are used as input when present (although in the majority of cases they are missing). It is not possible to use this assay bioactivity data as input to a RF approach, because the data is sparse and RF methods require complete input information. However, in Figure 3 we see that the neural network's peak accuracy increases to $R^2 = 0.71 \pm 0.03$ with 50 descriptors. We now achieve a significantly better quality of fit than RF (with one-tailed Welch's t-test $p = 3\times10^{-4}$) due to the strong bioactivity-bioactivity correlations present in the data. The neural network is able to successfully identify these stronger bioactivity-bioactivity relations, without them being swamped by the numerous but weaker descriptor-bioactivity correlations.

We particularly see the value of the bioactivity-bioactivity correlations with zero descriptors, where the neural network achieves $R^2 = 0.35 \pm 0.03$ due solely to bioactivity-bioactivity correlations. Random forest is not able to make predictions at all without any descriptors being present, as it cannot take the sparse bioactivity data as input, and so $R^2 = 0$. The ability to fit the data better than a leading QSAR method provides a solid platform for use of this neural network to impute assay bioactivity values.

3.2 Kinase data set

We now present a case study on the Kinase data set proposed as an exemplar for benchmarking predictive imputation methods,[13] as described in Section 2.1.

In this data set the validation data comprised the outliers from a clustering procedure, realistically representing the exploration of new chemical space. The best achieved coefficient of determination by a method in the literature is $R^2 = 0.434 \pm 0.009$ by the profile-QSAR 2.0 method,[13] which we re-implemented for this comparison. The DNN multi-target model discussed in Section 2.5 achieved $R^2 = 0.11 \pm 0.01$, the CMF method achieved $R^2 = -0.11 \pm 0.01$, and a conventional RF QSAR approach achieved only $R^2 = -0.19 \pm 0.01$, a result which is worse than random due to the extrapolation in chemical space required to reach the test set points.

Using our deep neural network we predict the assay bioactivity values and also the uncertainties in the predictions. With 100% of the predictions accepted, irrespective of the reported confidence, the neural network attains $R^2 = 0.445 \pm 0.007$, a significant improvement over the DNN, CMF, and RF approaches and similar to the profile-QSAR 2.0 method result. However, access to the uncertainties in the predictions gives us more knowledge about the neural network results. In particular, we can discard predictions carrying large uncertainty, and trust only those with smaller uncertainty. This

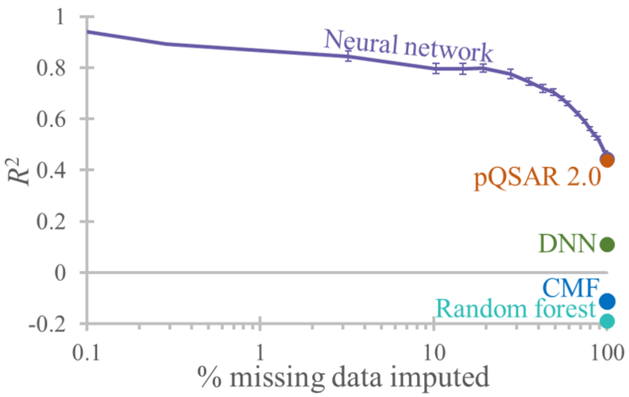

Figure 4: The coefficient of determination for predicting the activity of the clustered Kinase data set with percentage of data predicted. The cyan point is for the random forest approach, the blue point is the collective matrix factorisation (CMF) method, the dark green point is the deep neural network (DNN) approach, the orange point is the profile-QSAR 2.0 method, and the magenta line is the neural network proposed in this work. The magenta line shows that the accuracy of the neural network predictions increases when focussing on the most confident predictions, at the expense of imputing only a proportion of the missing data. This confirms that the reported confidences in the predictions correlate strongly with their accuracy. Error bars represent the standard error in the mean R^2 value over all 159 assays, and where not visible are smaller than the size of the points.

lets us focus on the most confident predictions only, at the expense of reporting fewer total predictions. When this is done, the quality of the remaining neural network predictions increases, as shown in Figure 4, demonstrating that the neural network is able to accurately and truthfully inform us about the uncertainties in its predictions; the confidence of predictions is correlated with their accuracy. The coefficient of determination reaches values of $R^2 > 0.9$, demonstrating effectively perfect predictions, when we complete only the most confident 1% of the data. We note that this focus on the most confident predictions, and corresponding increase in accuracy, is post-processing: only one model is trained, and the desired level of confidence can be specified and used to return only sufficiently accurate results.

The neural network is signifcantly more accurate than the DNN, CMF, and RF methods even when 100% of the predictions are accepted (with p-values 3×10^{-66}, 2×10^{-102}, and 2×10^{-107} respectively), and is significantly more accurate than pQSAR 2.0 when only the least confident 3% of predictions are discarded ($p = 3 \times 10^{-4}$). As shown in Figure 4, the accuracy improvement over the other methods increases substantially as a smaller fraction of the predictions are accepted.

The achieved $R^2 > 0.9$ exceeds the level of $R^2 = 0.7$ that is often taken as indicating accurate, reliable predictions in the presence of experimental uncertainty. In fact, the most confident 50% of the neural network's predictions all have $R^2 > 0.7$, permitting a nine-fold increase in the number of accurate predictions that can be used for further analysis, relative to the original sparse experimental measurements.

This high accuracy is achieved after approximately 120 core hours of training. The time to validate the data set is 0.1ms per compound for the neural network, versus 10ms per compound, 100 times longer, for the traditional random forest approach. This acceleration in generating predictions further enhances the real-world applicability of the neural network approach.

3.2.1 Analysis

It is informative to analyse the results that our neural network approach is able to calculate accurately, and compare this to preconceptions of how the algorithm functions. For example, as a data-driven approach, it might be assumed that the assays with the most training data would be most accurately predicted by the neural network. However, as shown in Figure 5, this is not the case; although the assay with least training data is that predicted least accurately, there is in general no correlation between the accuracy of the neural network's predictions and the amount of training data available to the algorithm. In particular, the two assays with most training data are relatively poorly captured by the neural network, with $R^2 < 0.2$ in both cases.

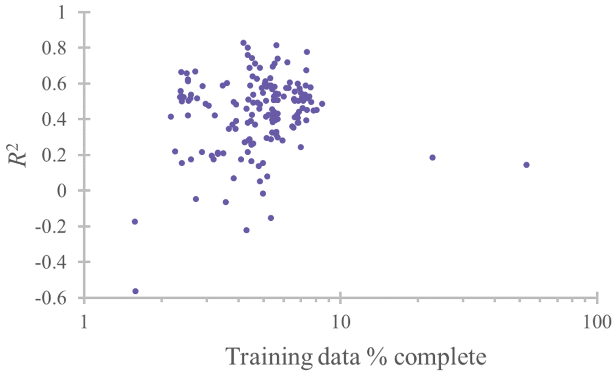

Figure 5: The coefficient of determination measured for each of the 159 kinase assays, plotted against the percentage of the data for that assay present in the training set.

Likewise, the most confident predictions are not for compounds 'closest' to those in the training set. The degree of separation can be measured in terms of the Euclidean distance between the points in the multi-dimensional space of descriptors used in the model. A representative example assay's data (ChEMBL assay 688660) is shown in Figure 6 where the training points (grey crosses) and test points (coloured points) are depicted in a 2-dimensional t-distributed stochastic neighbour embedding (t-SNE) generated using the StarDrop software package.[38] The levels of predictive confidence are fairly uniform with distance from the training data, confirming the algorithm's ability to confidently predict test points that are relatively far from the clusters of training points. In addition to this analysis, the Euclidean distance between every test point and its nearest neighbour training point was taken for all assays. This measure showed no correlation with the network's uncertainty or error, indicating that the neural network is operating beyond a nearest-neighbour approach in descriptor space, by exploiting assay-assay correlations that are carried across into assay-descriptor space.

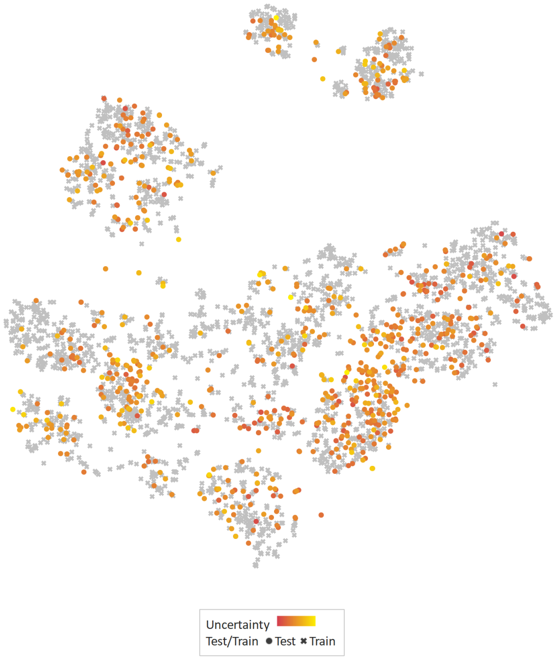

Figure 6: A 2-dimensional t-SNE embedding of the input descriptor space for ChEMBL assay 688660. The grey crosses show the training data and the coloured points show the test data with colour indicating the uncertainty estimate of the network in its predictions, where red indicates zero uncertainty and yellow a high uncertainty of 1 log unit.

3.2.2 Summary

We have shown that the neural network presented delivers similar quality predictions for assay bioactivity to the profile-QSAR 2.0 method when considering the full test set and that these methods outperform QSAR methods, including modern DNNs, and also outperforms matrix factorisation. In addition, a key advantage is that the neural network gives accurate uncertainties on its output, allowing us to prioritise only well-predicted assay activities, enabling an increase in the coefficient of determination for the predictions of the realistic data set from $R^2 = 0.445$ up to $R^2 > 0.9$ for a subset of the data. The ability to tune accuracy with amount of data predicted is an invaluable tool for scientists, fueling confidence in results and permitting a focus on only high-quality predictions. These most confident predictions are also not for the most complete assays or the most similar test points to the training data, showing that the neural network approach is able to learn more complex and powerful representations of the assay bioactivity data.

4 Conclusions

We have presented a new neural network imputation technique for predicting bioactivity, which can learn from incomplete bioactivity data to improve the quality of predictions by using correlations between both different bioactivity assays, and also between molecular descriptors and bioactivities. This results in a significant improvement in the accuracy of prediction over conventional QSAR models, even those using modern deep learning methods, particularly for challenging data sets representing an extrapolation to new compounds that are not well represented by the set used to train the model. This is representative of many chemistry optimisation projects which, by definition, explore new chemical space as the project proceeds.

The method presented can also accurately estimate the confidence in each individual prediction, enabling attention to be focussed on only the most accurate results. It is important to base decisions in a discovery project on reliable results to avoid wasted effort pursuing incorrectly selected compounds or missing opportunities by inappropriately discarding potentially valuable compounds.[39] On the Kinase example data set, we demonstrated that 50% of the missing data could be filled in with $R^2 > 0.7$, which is considered to represent a high level of fidelity between prediction and experiment.

The ability to make simultaneous, accurate predictions across multiple assays will lend itself well to the problem of selectivity across multiple targets.[40,41] The method is general, so can apply beyond the binding assay data used in this analysis, for example to direct or downstream functional assays; and the method can even make accurate predictions beyond pIC_{50} values, including physicochemical, absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties. Therefore, it has a broad application for identification of additional active compounds within a database, recognition of the most influential chemical properties, prediction of selectivity profiles, and the selection of compounds for progression to downstream ADMET assays.

Conflict of Interest Statement Matthew Segall, Peter Hunt and Ben Irwin are employees of Optibrium Limited, which develops the StarDrop$^{\mathrm{TM}}$ software used herein for analysis of the results. Thomas Whitehead and Gareth Conduit are employees of Intellegens Limited, which develops the Alchemite$^{\mathrm{TM}}$ software used in this work.

Acknowledgement Gareth Conduit acknowledges the financial support of the Royal Society and Gonville & Caius College. There is Open Access to this paper and data available at `https://www.openaccess.cam.ac.uk`.

Supporting Information Available

The following files are available free of charge:

- Adrenergic_dataset.csv: Dataset used in Section 3.1
- Kinase_training_w_descriptors.csv: Training dataset used in Section 3.2
- Kinase_test_w_descriptors.csv: Test dataset used in Section 3.2

This information is available free of charge via the Internet at `http://pubs.acs.org`.

References

(1) Gaulton, A.; Bellis, L.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Akhtar, R.; Bento, A.; Al-Lazikani, B.; Michalovich, D.; Overington, J. ChEMBL: A Large-scale Bioactivity Database For Chemical Biology and Drug Discovery. *Nucleic Acids Res.* **2012**, *40*, 1100.

(2) Bento, A.; Gaulton, A.; Hersey, A.; Bellis, L.; Chambers, J.; Davies, M.; Krüger, F.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. The ChEMBL Bioactivity Database: an Update. *Nucleic Acids Res.* **2014**, *42*, 1083.

(3) Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; Consonni, V.; Kuz'min, V. E.; Cramer, R.; Benigni, R.; Yang, C.; Rathman, J.; Terfloth, L.; Gasteiger, J.; Richard, A.; Tropsha, A. QSAR Modeling: Where Have You Been? Where Are You Going To? *J. Med. Chem.* **2014**, *57*, 4977–5010.

(4) Wold, S.; Sjostrom, M.; Eriksson, L. In *The Encyclopedia of Computational Chemistry*; Schleyer, P., Allinger, N., Clark, T., Gasteiger, J., Kollman, P., Schaefer III, H., P., S., Eds.; Chichester, UK: John Wiley and Sons, 1999; pp 1–16.

(5) Gao, C.; Cahya, S.; Nicolaou, C.; Wang, J.; Watson, I.; Cummins, D.; Iversen, P.; Vieth, M. Selectivity Data: Assessment, Predictions, Concordance, and Implications. *J. Med. Chem.* **2013**, *56*, 6991.

(6) Schurer, S. C.; Muskal, S. M. Kinome-wide Activity Modeling from Diverse Public High-quality Data Sets. *J. Chem. Inf. Model.* **2013**, *53*, 27.

(7) Christmann-Franck, S.; van Westen, G.; Papadatos, G.; Beltran Escudie, F.; Roberts, A.; Overington, J.; Domine, D. Unprecedently Large-Scale Kinase Inhibitor Set Enabling the Accurate Prediction of Compound Kinase Activities: A Way toward Selective Promiscuity by Design? *J. Chem. Inf. Model.* **2016**, *56*, 1654.

(8) Subramanian, V.; Prusis, P.; Xhaard, H.; Wohlfahrt, G. Predictive Proteochemometric Models for Kinases Derived from 3D Protein Field Based Descriptors. *MedChemComm* **2016**, *7*, 1007.

(9) Merget, B.; Turk, S.; Eid, S.; Rippmann, F.; Fulle, S. Profiling Prediction of Kinase Inhibitors: Toward the Virtual Assay. *J. Med. Chem.* **2017**, *60*, 474.

(10) Doucet, J.; Xia, H.; Panaye, A.; Fan, B. Nonlinear SVM Approaches to QSPR/QSAR Studies and Drug Design. *Curr. Comput.-Aided Drug Des.* **2007**, *3*, 263–289.

(11) Obrezanova, O.; Csanyi, G.; Gola, J.; Segall, M. Gaussian Processes: a Method for Automatic QSAR Modeling of ADME Properties. *J. Chem. Inf. Model.* **2007**, *47*, 1847–1857.

(12) Martin, E.; Mukherjee, P.; Sullivan, D.; Jansen, J. Profile-QSAR: a Novel Meta-QSAR Method that Combines Activities Across the Kinase Family to Accurately Predict Affinity, Selectivity, and Cellular Activity. *J. Chem. Inf. Model.* **2011**, *51*, 1942.

(13) Martin, E.; Valery R. Polyakov, V.; Tian, L.; Perez, R. Profile-QSAR 2.0: Kinase Virtual Screening Accuracy Comparable to Four-Concentration IC50s for Realistically Novel Compounds. *J. Chem. Inf. Model.* **2017**, *57*, 2077.

(14) Lo, Y.-C.; Rensi, S. E.; Torng, W.; Altman, R. B. Machine Learning in Chemoinformatics and Drug Discovery. *Drug Discovery Today* **2018**, *23*, 1538 – 1546.

(15) Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The Rise of Deep Learning in Drug Discovery. *Drug Discovery Today* **2018**, *23*, 1241 – 1250.

(16) Mayr, A.; Klambauer, G.; Unterthiner, T.; Steijaert, M.; Wegner, J. K.; Ceulemans, H.; Clevert, D.-A.; Hochreiter, S. Large-scale Comparison of Machine Learning Methods for Drug Target Prediction on ChEMBL. *Chemical Science* **2018**, *9*, 5441.

(17) Popova, M.; Isayev, O.; Tropsha, A. Deep reinforcement learning for de novo drug design. *Science Advances* **2018**, *4*, eaap7885.

(18) Conduit, B.; Jones, N.; Stone, H.; Conduit, G. Design of a Nickel-base Superalloy with a Neural Network. *Materials and Design* **2017**, *131*, 358.

(19) Conduit, B.; Jones, N.; Stone, H.; Conduit, G. Probabilistic Design of a Molybdenum-base Alloy using a Neural Network. *Scripta Materialia* **2018**, *146*, 82.

(20) Verpoort, P.; MacDonald, P.; Conduit, G. Materials Data Validation and Imputation with an Artificial Neural Network. *Computational Materials Science* **2018**, *147*, 176.

(21) Ertl, P.; Rhodes, B.; Slezer, P. Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its application to the Prediction of Drug Transport. *J. Med. Chem.* **2000**, *43*, 3714–3717.

(22) Abraham, M.; McGowan, J. TheUuse of Characteristic Volumes to Measure Cavity Terms in Reversed-phase Liquid-chromatography. *Chromatographia* **1987**, *23*, 243–246.

(23) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1998**, *28*, 31–36.

(24) Heskes, T. Practical Confidence and Prediction Intervals. Advances in Neural Information Processing Systems 9. 1997; pp 176–182.

(25) Papadopoulos, G.; Edwards, P.; Murray, A. Confidence Estimation Methods for Neural Networks: a Practical Comparison. *IEEE Transactions on Neural Networks* **2001**, *12*, 1278.

(26) Krishnan, T.; McLachlan, G. *The EM Algorithm and Extensions*; Wiley, 2008.

(27) He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Patter Recognition (CVPR). 2016; pp 770–778.

(28) Mahfoud, S.; Goldberg, D. Parallel Recombinative Simulated Annealing: A Genetic Algorithm. *Parallel Computing* **1995**, *21*, 1.

(29) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

(30) Xu, Y.; Ma, J.; Liaw, A.; Sheridan, R. P.; Svetnik, V. Demystifying Multitask Deep Neural Networks for Quantitative Structure-Activity Relationships. *J. Chem. Inf. Model.* **2017**, *57*, 2490–2504.

(31) Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Goodfellow, I.; Harp, A.; Irving, G.; Isard, M.; Jia, Y.;

Jozefowicz, R.; Kaiser, L.; Kudlur, M.; Levenberg, J.; Mané, D.; Monga, R.; Moore, S.; Murray, D.; Olah, C.; Schuster, M.; Shlens, J.; Steiner, B.; Sutskever, I.; Talwar, K.; Tucker, P.; Vanhoucke, V.; Vasudevan, V.; Viégas, F.; Vinyals, O.; Warden, P.; Wattenberg, M.; Wicke, M.; Yu, Y.; Zheng, X. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. 2015; `https://www.tensorflow.org/`, Software available from tensorflow.org.

(32) Nesterov, Y. A Method of Solving a Convex Programming Problem with Convergence Rate $O(1/k^2)$. *Soviet Mathematics Doklady* **1983**, 372–376.

(33) Mehta, R.; Rana, K. A review on matrix factorization techniques in recommender systems. 2017 2nd International Conference on Communication Systems, Computing and IT Applications (CSCITA). 2017; pp 269–274.

(34) Töscher, A.; Jahrer, M.; Bell, R. M. The BigChaos Solution to the Netflix Grand Prize. 2009.

(35) Singh, A. P.; Gordon, G. J. Relational Learning via Collective Matrix Factorization. Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA, 2008; pp 650–658.

(36) Cortes, D. Cold-start recommendations in Collective Matrix Factorization. *CoRR* **2018**, *abs/1809.00366*.

(37) Žuvela, P.; Liu, J. J.; Macur, K.; Bączek, T. Molecular Descriptor Subset Selection in Theoretical Peptide Quantitative Structure-Retention Relationship Model Development Using Nature-Inspired Optimization Algorithms. *Anal. Chem.* **2015**, *87*, 9876–9883.

(38) StarDrop. `https://www.optibrium.com/stardrop/`, Accessed: 2018-10-26.

(39) Segall, M.; Champness, E. The Challenges of Making Decisions using Uncertain Data. *J. Comput.-Aided Mol. Des.* **2015**, *29*, 809–816.

(40) Sciabola, S.; Stanton, R. V.; Wittkopp, S.; Wildman, S.; Moshinsky, D.; Potluri, S.; Xi, H. Predicting Kinase Selectivity Profiles Using Free-Wilson QSAR Analysis. *Journal of Chemical Information and Modeling* **2008**, *48*, 1851–1867, PMID: 18717582.

(41) Kothiwale, S.; Borza, C.; Pozzi, A.; Meiler, J. Quantitative Structure–Activity Relationship Modeling of Kinase Selectivity Profiles. *Molecules* **2017**, *22*.